

基于多元线性回归模型的房价影响因素探究

黄予暄*, 叶梓晴, 樊喆

暨南大学 暨南大学伯明翰大学联合学院, 广东 广州 511443

DOI:10.61369/ASDS.2026050014

摘 要 : 房价波动不仅关系居民的居住福利, 更是宏观经济运行与市场风险评估的重要指标。探究房价内部驱动机制能为购房群体决策、房企定价谋划及行政部门调控提供量化参考。本研究以2014至2015年间美国华盛顿州金县15183条房产交易记录作为实证样本, 结合特征价格理论搭建多元线性回归估算框架, 旨在剖析各类物理属性及区位条件对房屋价值的具体作用程度。整体评估过程涵盖数据探索性剖析、多重共线与异方差排查、残差诊断等统计检验, 配合逐步回归算法完成特征变量筛选。鉴于样本呈现异方差特性, 后续估算直接引入稳健标准误以构建更为严密的修正路径。数据实证反馈, 居住面积扩张、海景资源占有、视野开阔度提升、建筑成色优化及设计水准拔高均对资产溢价具备强效的正向驱动, 环境景观要素已深度融入资本定价体系。剥离总面积规模干扰项后, 卧室分隔数量增多及医疗设施通勤距离拉长反而对资产估值形成显著的负向制约。上述发现清晰地刻画出金县房产交易板块的核心定价法则, 足以支撑多元市场参与主体及管理机构的科学决策。

关 键 词 : 多元回归分析; 逐步回归; 多重共线性检验; 残差检验; 异方差检验

An Exploration of Factors Influencing Housing Prices Based on a Multiple Linear Regression Model

Huang Yuxuan*, Ye Ziqing, Fan Zhe

Jinan University - University of Birmingham Joint Institute at Jinan University, Jinan University, Guangzhou, Guangdong 511443

Abstract : Housing price volatility affects not only the residential welfare of the public but also serves as a crucial indicator for macroeconomic performance and market risk assessment. Investigating the intrinsic driving mechanisms of housing prices provides a quantitative reference for homebuyers' decision-making, developers' pricing strategies, and government regulatory interventions. Utilizing an empirical sample of 15,183 property transaction records from King County, Washington, USA, between 2014 and 2015, this study constructs a multiple linear regression framework based on the hedonic price theory to dissect the specific impacts of various physical attributes and locational conditions on property values. The comprehensive evaluation process encompasses statistical procedures including exploratory data analysis, multicollinearity and heteroskedasticity testing, and residual diagnostics, complemented by a stepwise regression algorithm for feature selection. Given the heteroskedastic nature of the sample, robust standard errors are explicitly incorporated into subsequent estimations to establish a more rigorous analytical model. Empirical findings indicate that the expansion of living area, the presence of a waterfront view, enhanced visual outlooks, superior building conditions, and elevated architectural design levels all exert a robust positive driving force on property premiums, demonstrating that environmental and landscape elements are deeply capitalized into the pricing system. After controlling for the confounding variable of total property size, an increased number of bedrooms and a longer commuting distance to medical facilities impose a significant negative constraint on property valuation. These findings clearly delineate the core pricing principles within the King County real estate market, offering robust support for the evidence-based decision-making of diverse market participants and regulatory authorities.

Keywords : multiple regression analysis; stepwise regression; multicollinearity test; residual analysis; heteroscedasticity test

基金项目: 2025年度广东省大学生创新创业训练计划支持项目“复杂数据下自相关结构误差项参数对指向性拟合优度检验表现的敏感性分析”(项目编号: CX25543)。

作者简介:

叶梓晴, 暨南大学暨南大学伯明翰大学联合学院, 本科在读, 研究方向为经济统计学;

樊喆, 暨南大学暨南大学伯明翰大学联合学院, 本科在读, 研究方向为经济统计学。

通讯作者: 黄予暄, 暨南大学暨南大学伯明翰大学联合学院, 本科在读, 研究方向为经济统计学。

引言

房地产行业属于国民经济支柱产业，同人民居住福利、宏观经济金融风险控制存在联系。美国华盛顿州金县包含西雅图等主要都市圈，这里因为商业活动频繁、人口众多而呈现出典型的房产交易景象。目前学术团队和实务部门一般依靠数据科学配合计量经济学来建立统计模型，目的是准确地量房屋的物理属性、区位特点以及周边配套对房价的影响。

2014年5月至2015年5月，从金县的15183份房地产交易记录中提取出来做数据支持，建立多元线性回归模型来分析房价内部的主要影响因素以及作用机制。整体研究路径按照“数据预处理－描述性统计－模型构建－实证检验与修正－结论分析”的逻辑结构来展开，实证部分用普通最小二乘法（OLS）和逐步回归法做特征选取。对于截面回归中经常出现的异方差和多重共线性问题，本文用对数变换、标准化以及 White 稳健标准误来加以修正。该分析框架准确地确定了居住面积、建筑水平等主要变量实际所起的作用，最后得到一个具有很强解释力的稳健模型。

从实践角度来讲，本文所测算出的结论可以为购房者提供客观的房产溢价空间的估计，也可以成为房企选址定价、行政机构市场监测的量化标准。结合目前的房地产数字化实务，简单的高效多元线性回归模型仍然是发掘房产数据潜在价值、快速产生有效参考价的主要底层算法，从而更加强化本文所建模型的直接使用价值^[1]。本文末尾依靠各个核心变量回归系数的经济学解释，对金县房地产市场运行机理进行全方位的剖析。

一、数据与变量预处理

（一）数据来源

本研究使用的数据来自于 Kaggle 上的数据集 House Sales in King County, USA，其中包括金县从2014年5月到2015年5月的房屋销售数据，原始数据共15183条。

（二）变量选取与定义

为了全面探究房价的影响因素，本文从房屋物理属性、外部景观、区位可达性这三个方面选取了10个自变量。各个变量的定义和赋值如下所示：

因变量：销售价格

物理属性自变量为卧室数、浴室数、居住面积、楼层数、有无地下室、建筑设计水平、成色。

外部景观自变量：包括是否海景房、视野。

区位自变量是距医院的距离，属于医疗配套可达性。

以下是变量定义：

表1：变量选取与详细定义 Table 1: Selecting & Defining Variables		
	变量名	详细定义
因变量	销售价格	房屋成交时的市场价格
	卧室数	房产的卧室数量
	浴室数	房产的浴室数量
	居住面积	房产的室内面积（地上加地下）
	楼层数	房产的楼层数
	是否海景房	房产是否是海景房
	视野	房产的视野等级，即房周围的景色是否好看
自变量	成色	房产的成色，即“几成新”
	建筑设计水平	房产的建筑设计水平，由专家审定，评判标准包括如房间结构的划分、窗户的朝向等是否合理
	距医院的交通	计算每间房产距华盛顿大学医学中心的距离
	是否有地下室	地下建筑面积为 0 记为“无”，非 0 则记为“有”

（三）数据预处理

为了满足多元线性回归模型的基本假设，保证参数估计的稳健性以及有效性，本文对原始数据进行了预处理。统计学常规的预处理手段主要有数据清洗和数据归约两种。对于数据清洗过程中的主要工作有异常检测、剔除近似重复的对象和修补缺失的数据等，以保证总体分布的正常，并且得到最终的预测结果^[2]。

1. 缺失值处理

对原始数据进行完整性检验，删除掉有空值的样本。由于缺失部分占比较低，此操作可以保证观测序列的连续性，并不会给总体统计特征带来太大的偏差。

2. 异常值处理

对于连续型自变量，用四分位距 (IQR) 法来识别和修正。具体计算利用变量的第一四分位数 (Q1) 与第三四分位数 (Q3) 定义 $IQR=Q3-Q1$ ，超出 $[Q1-1.5*IQR, Q3+1.5*IQR]$ 区间的观测值即判定为统计异常。为了保证样本量以及减少噪声干扰，把这类异常值全部替换成对应的变量的中位数。对于虚拟变量“是否海景房”，本文使用逻辑一致性检验来保证它只取值于 {0, 1}，如果有不符合的数值就将其设为1或者0。

3. 变量变换

因变量“销售价格”数值横跨数十万到百万美元量级，本文对它进行对数变换，即取 $\log(Price+1)$ 。内部加1设定旨在规避零值取对数引发计算错误。对数变换可以改变数据的大小，减小偏斜度，使序列接近正态分布假设、稳定的方差、降低模型对于极端数值的敏感性。

4. 数据标准化

为了消除自变量的物理量纲干扰，提高回归参数的可比性以及模型整体的稳定性，对于居住面积、建筑设计水准、卫浴和卧室数量、房屋成色、楼层总数和视野评分这些连续型变量统一使用 Z-score 标准化，即将转换序列均值设为0，标准差设为1。就

“距医院交通距离”这个特有变量而言，观测数值大多处在0.1的附近，同居住面积这些高量级指标相比，存在着明显的差别，因此研究小组采取了归一化（Min-Max Scaling）的办法，将此指标调整到[0, 1]的闭区间之内。此项处理可以保持指标原始的差别特点，而且提高模型对跨量级数据的处理速度。

二、探索性数据分析（EDA）

（一）连续变量分布特征

建立回归模型第一步就是筛选数据基础特征。原始数据里隐藏的不一致或者极端分布，如果不用探查就直接进入模型，很容易导致最终回归结果出现偏差^[3]。本文采用直方图观察各个主要变量的分布形状，发现样本的集中趋势及偏态。

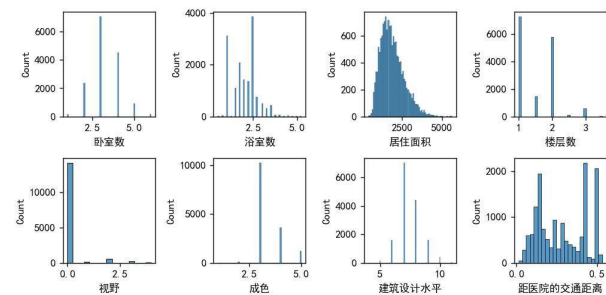


图1：连续型自变量分布直方图

Figure 1: Histogram of Continuous Independent Variables Distribution

就房屋的物理属性而言，卧室的数量有明显的集中现象，大多数房产有2-3间，3居室占主导地位。浴室数量分布较为多样，大部分集中在1~2.5个之间。居住面积分布具有典型的右偏态（正偏）特征。统计数据表明绝大多数房产室内面积处于较小的数值区间内，只有极少数高端房产有超大的建筑规模。长尾分布特征正好证明后面将对因变量和核心面积指标做对数变换、标准化处理的必要性。

楼层数和建筑设计水平的分布呈分段集中状。楼层数大多为1层或者2层，符合金县居住建筑的普遍形态。建筑设计水平处于7到8级，客观上体现样本房产整体具备中等偏上的质量标准。

根据探究区位因素可以发现，距医院的交通距离整体上比较分散，但0.1-0.2公里和0.4公里周边的频数有明显高峰，说明样本房产医疗配套资源的空间分布是不一样的。

（二）离散变量构成分析

在完成了房屋物理属性分析之后，本文又用饼图对“是否海景房”和“有无地下室”这两个重要的离散变量的构成比例进行了量化。

统计数据说明样本房产海景房比例仅为0.1%，这样一种明显的失衡分布情况意味着金县的房地产市场里海景资源十分缺乏，这将给之后的回归分析赋予很强的溢价效果。根据探查结果可知，拥有地下室房产的占总房产数的38.1%，而无地下室房产的占总房产数的61.9%。变量交互分布结果表明，市场中绝大多数的房产都没有海景和地下室两个缺失情况。以上离散变量分布态势直接体现金县住房市场的总体结构，也为后面用虚拟变量做分类研

究提供数据基础。

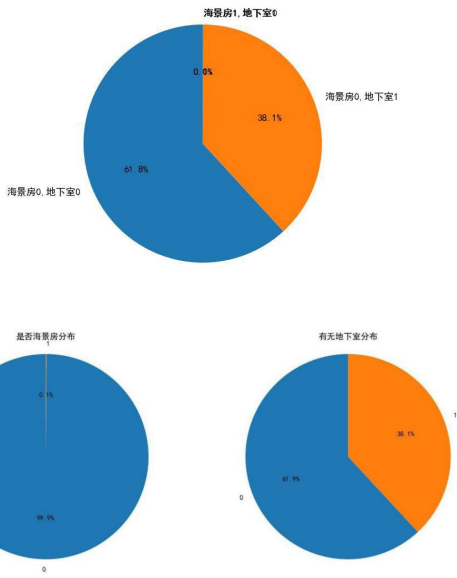


图2：离散型自变量构成比例饼图

Figure 2: Pie Chart of Proportion for Discrete Independent Variables

（三）变量间相关性分析

为了进一步找出各个变量之间的线性相关程度以及是否存在多重共线性问题，本文计算了变量间 Pearson 相关系数，并画出了相关系数矩阵热力图。

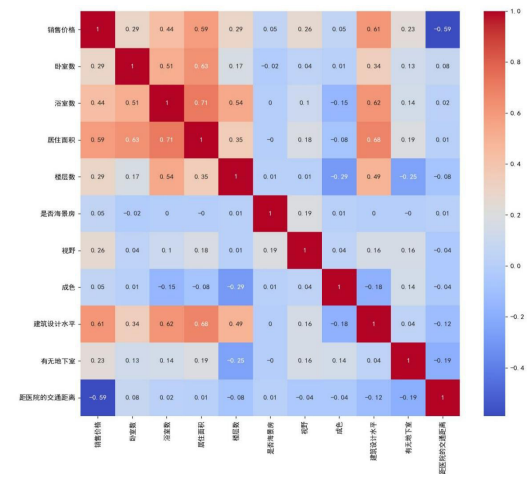


图3：变量相关系数矩阵热力图

Figure 3: heatmap of variable correlation coefficient matrix

由统计结果可知，销售价格同很多自变量有显著的正相关。居住面积与销售价格的相关系数为0.59，建筑设计水平、浴室数的系数分别为0.61、0.44，初步证明房屋物理规格构成是决定金县房价的主要因素。交通距离和售价呈负相关，地理位置对于房产定价有着较大的影响。

探究自变量内部相关性可以发现热力图显示部分物理属性指标有较强的共线性倾向。居住面积、浴室数和卧室数的皮尔逊相关系数分别为0.71、0.63，客观上说明大户型房产一般会配备更多的房间，建筑水平和居住面积也存在较高的相关性（0.68）。

部分自变量虽然有中等程度以上的相关性，但是初步的观察结果表明，各项系数都没有超过0.8的警戒线。之后会使用方差膨胀因子 (VIF) 做严密的多重共线性检验，保证底层回归估计的正确性。

三、模型估计与实证检验

(一) 全变量初步回归分析

本文用最小二乘法 (OLS) 对全部10个自变量的模型进行初步回归估计，回归结果如下所示。

表2: 全变量初步回归结果

Table 2: Preliminary Regression Results All Variables						
变量名称	coef	std err	t	P> t	0.025	0.975
const	13.4284	0.005	2925.612	0	13.419	13.437
卧室数	-0.0087	0.002	-3.582	0	-0.013	-0.004
浴室数	0.003	0.003	0.988	0.323	-0.003	0.009
居住面积	0.183	0.003	54.179	0	0.176	0.19
楼层数	-0.0034	0.002	-1.361	0.173	-0.008	0.001
是否海景房	0.4731	0.059	8.023	0	0.358	0.589
视野	0.0509	0.002	26.58	0	0.047	0.055
成色	0.0456	0.002	23.8	0	0.042	0.049
建筑设计水平	0.1311	0.003	47.219	0	0.126	0.137
有无地下室	0.0051	0.004	1.179	0.238	-0.003	0.014
距医院的交通距离	-0.8866	0.007	-131.869	0	-0.9	-0.873

统计量名称	数值
R-squared	0.764
Adj. R-squared	0.764
F-statistic	4925
Prob (F-statistic)	0
Log-Likelihood	1190.8
AIC	-2360
BIC	-2276
Omnibus	259.943
Prob(Omnibus)	0
Skew	0.013
Kurtosis	3.913
Durbin-Watson	1.992
Jarque-Bera (JB)	528.194
Prob(JB)	2.01e-115
Cond. No.	57.5

统计数据表明模型 R^2 和调整后的 R^2 都是0.764，说明本模型可以解释销售价格大约76.4%的变化。F 统计量为4925并且对应的 P 值小于0.01，说明回归方程整体上是显著的。

探究变量显著性可知回归系数有明显的差异性。房屋的物理规格、景观配套等各方面因素都会对房价产生显著的正向影响。居住面积 (coef=0.1830, $P<0.01$)、建筑设计水平 (coef=0.1311, $P<0.01$)、视野评分 (coef=0.0509, $P<0.01$) 及房屋成色 (coef=0.0456, $P<0.01$) 均顺利通过1% 显著性水平检验。“是否海景房”虚拟变量系数为0.4731, $P < 0.01$ ，并且直

观地显示了稀缺的景观资源所具有的很强的溢价属性。

从区位要素的分析可知，核心变量有很强的负向制约作用。“距医院的交通距离”的系数为负的，说明房产销售价格随它离医疗中心距离的增加而减少。

有些观测变量第一次测算没有达到预期的显著性。卧室数 (coef=-0.0087, $P<0.01$) 具备统计学意义却呈现负向干扰属性。浴室数、楼层数、是否有地下室三个指标的 P 值全部大于0.05显著性水平，说明全变量拟合框架中以上因素对销售价格的边际贡献非常小。

观察基本假设检验结果可知，Durbin-Watson 统计量为1.992，与临界值 2 比较接近，说明残差序列不存在明显的自相关性。全变量模型有较强的解释效能，内部夹杂的不显著变量和潜在的异方差问题之后的研究要直接进行架构精简，并且需要使用稳健性参数进行修正。

(二) 模型诊断与统计检验

在得到初步回归估计之后，对模型的基本假设进行严格的统计诊断，保证回归推断的科学性、有效性。

为了防止自变量之间存在多重共线性，用方差膨胀因子 (VIF) 来量化的检验。方差膨胀因子是解释变量夹杂多重共线性相对于无共线性状态下的方差比值，VIF 数值增大直观反映共线性变大。经验法则规定 $0 < VIF < 10$ 为没有多重共线性的区间， $10 \leq VIF < 100$ 为存在共线性干扰的区间， $VIF \geq 100$ 为严重的共线性状态^[4]。

表3: 全变量模型方差膨胀因子 (VIF) 检验结果

Table 3: Variance inflation factor(VIF) test result of the full variable model	
变量	VIF
const	6.386470905021658
卧室数	1.7893026497069366
浴室数	2.744949264549292
居住面积	3.4578771443893825
楼层数	1.8899434188653077
是否海景房	1.0406427740985638
视野	1.1103674714811222
成色	1.1147991996666844
建筑设计水平	2.3354331230186376
有无地下室	1.3309880134478493
距医院的交通距离	1.123095130097447

从测算结果可以得知各个预测变量的 VIF 值都在1.04到3.46之间，远远小于10。居住面积 VIF 冲顶到3.46，符合它和卧室、浴室数量的中度相关性，并且数值仍在安全范围内。整体检验发现自变量之间没有严重的多重共线性，并且回归系数的估计比较稳定。

面对残差方差齐性诊断，用 Breusch-Pagan (BP) 检验法来检查。

表4: 全变量模型残差异方差 (BP) 检验结果

Table 4: Residual heteroscedasticity (BP) Test Results of the Full Variable Model		
检验统计量	值	p 值
检验统计量	278.6524	0.0000
F 检验统计量	28.3656	0.0000

测算输出 BP 统计量高达278.6524且对应的 P 值显著小于

0.05 ($P=0.0000$), F 检验统计量 (28.3656) 的 P 值同样为 0.0000。极值表现推翻了残差项同方差性的原假设, 说明模型内部存在明显的异方差性。该种现象表明普通最小二乘法的标准误估计有误, 需要用稳健标准误 (Robust Standard Errors) 来修正。

对于残差分布的正态性假设, 用残差散点图和分位数 - 分位数图 (Q-Q 图) 来检验,

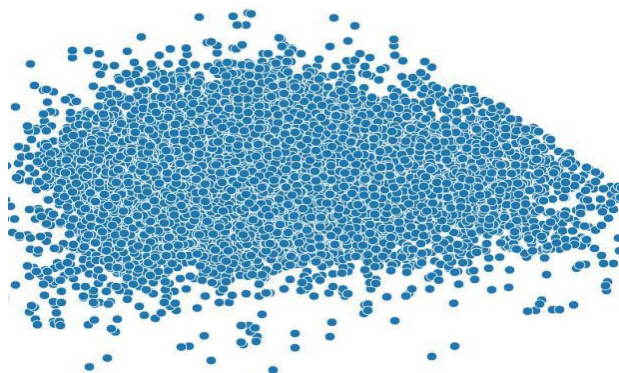


图4: 全变量模型残差与拟合值散点图

Figure 4: Scatterplot of residual vs. fit from full variable model

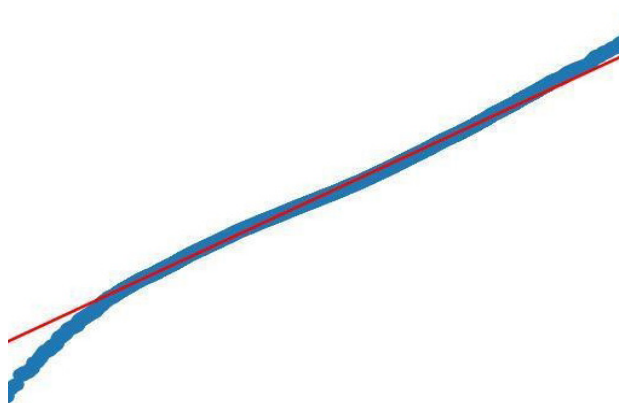


图5: 全变量模型残差正态 Q-Q 图

Figure 5: Normal Q-Q plot of the residuals from the full variable model

从残差与拟合值的散点图中可以看出, 残差基本处在零线左右随机分布的状态, 没有出现明显结构化的特点, 模型的线性设定逻辑是合理的。

残差正态 Q-Q 图中大部分观测点都接近对角参考线, 说明残差大致服从正态分布。

四、模型稳健性分析与优化

确定最后的多变量解释框架之前, 必须对模型的基本设定进行严格的甄别, 避免出现模型设定偏差的情况, 保证特征选择是客观合理的。本部分将按照非线性特征诊断、基于机器学习算法的特征双重验证的顺序来建立一个具有较好统计性质的修正模型。

(一) 模型函数形式甄别

经典特征价格模型 (Hedonic Price Model) 实证框架中, 因

变量和自变量之间的映射关系不一定是一条严格的线性组合。本研究主要研究金县房价和核心物理、区位特征的真实影响路径, 并判断是否存在边际效用递减或者阈值效应等高阶非线性特征, 对保留下来的连续型自变量和对数变换后的因变量 ($\ln(\text{Price})$) 进行视觉诊断和函数形式甄别, 同时输出各个核心变量 (居住面积、建筑设计水平、距医院交通距离等) 与对数房价散点图组

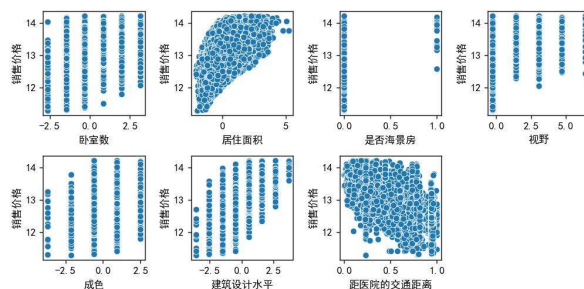


图6: 核心自变量与对数销售价格散点图

Figure 6: Scatter diagram for the core independent variable vs. Logarithm of sales price

从汇总散点图的输出结果可知, 核心解释变量 (尤其是居住面积、建筑设计水平等) 和对数房价之间存在明显的、稳定的单调递增关系, 数据点大部分都在一条线性或者接近线性的轨迹上。所有的观测散点群都没有显示出明显的 U 型或者倒 U 型的拐点, 客观上说明本样本区间内面积过大导致单价边际递减的极端非线性结构是不存在的。

根据以上诊断, 按照计量经济学建模的奥卡姆剃刀原则, 研究决定采用纯线性方程 (对数 - 线性形式) 的基础设定, 暂时不加入平方项或者高阶交互项。该决策的合理性是以没有强有力的非线性先验理论支持和直观的视觉证据为前提的, 强行加入高阶项会减少模型自由度, 增大过拟合的风险。同时线性设定可以保证模型简洁, 使变量系数直接解释为半弹性, 保证实证结果有经济学解释力以及政策指导意义。

(二) 逐步回归分析

为了进一步优化并检验前面建立的修正模型, 本文用逐步回归的方法选择出最具有预测能力的特征变量。逐步回归属于常规特征选择手段, 依靠迭代添加或者剔除特征来搭建模型提升预测效能, 削减过拟合危险。

借鉴已有房价实证分析修正经验, 对于初始全模型夹杂冗余或者不显著的变量, 采用逐步回归客观剔除干扰因素, 得到最优变量组合的方法来提高模型拟合优度和统计稳健性是标准的操作方式^[6]。分析具体变量筛选环节, 逐步回归法一般会设置引入变量的显著性水平临界值 (F -to-enter) 和剔除变量的显著性水平临界值 (F -to-remove), 全程用 F 检验来判断变量是否需要加入或者剔除, 直到回归方程内部没有新的变量可以加入、也没有旧的变量需要剔除为止^[6]。本次实验阶段逐步回归选择器用前向选择的方式开始, 从空模型出发, 按照整体性能提升最明显的原则逐渐添加特征。选择以负均方误差 (MSE) 为评价标准, 用 5 折交叉验证 (5-fold CV) 来检验模型是否稳定、泛化能力强。

表5：逐步回归模型估计结果						
Table 5: The estimation result of the stepwise regression model.						
变量名称	coef	std err	z	P> z	0.025	0.975
const	13.4308	0.004	3519.263	0	13.423	13.438
居住面积	0.1849	0.003	59.813	0	0.179	0.191
建筑设计水平	0.1299	0.003	50.151	0	0.125	0.135
是否海景房	0.4697	0.059	7.969	0	0.354	0.585
卧室数	-0.0083	0.002	-3.448	0.001	-0.013	-0.004
成色	0.0464	0.002	24.881	0	0.043	0.05
距医院的交通距离	-0.8874	0.006	-137.052	0	-0.9	-0.875
视野	0.0513	0.002	27.069	0	0.048	0.055

统计量名称	数值
Omnibus	261.535
Prob(Omnibus)	0
Skew	0.011
Kurtosis	3.918
Durbin-Watson	1.992
Jarque-Bera (JB)	532.933
Prob(JB)	1.88e-116
Cond. No.	47.9

逐步回归筛选结果得到最终入选模型的特征有居住面积、建筑设计水平、是否海景房、卧室数、房屋成色、距医院交通距离、视野评分。该模型有7个特征变量，F统计量达到7033，R²值一直为0.764。测算环节证明，逐步回归自动筛选出的特征变量和前面依靠统计显著性诊断保留下来的变量完全吻合。该流程可以实现算法驱动和统计推断的双重验证，直接证明修正模型变量选择的合理性，并给总体研究结论提供更加科学系统的理论基础。

（三）修正模型诊断

初步全变量回归分析表明，浴室数、楼层数、地下室是否具有三个变量的系数都不满足统计显著性检验的要求（P值均大于0.05），说明这三个变量对模型的解释能力比较小。本研究直接去掉以上变量来提高模型的精简程度和预测能力。

为了解决原模型的异方差问题，修正模型用稳健标准误（Robust Standard Errors，HCO 协方差类型）来消除对参数估计可靠性的干扰。该路径用改变标准误计算方法来减少模型异方差的敏感性，得到更加稳健的估计结果。

表6：稳健标准误修正模型回归结果						
Table 6: Regression Results of Robust Standard Error Correction Model						
变量名称	coef	std err	z	P> z	0.025	0.975
const	13.4308	0.004	3602.51	0	13.424	13.438
卧室数	-0.0083	0.003	-3.248	0.001	-0.013	-0.003
居住面积	0.1849	0.003	57.716	0	0.179	0.191
是否海景房	0.4697	0.048	9.766	0	0.375	0.564
视野	0.0513	0.002	24.066	0	0.047	0.056
成色	0.0464	0.002	23.962	0	0.043	0.05
建筑设计水平	0.1299	0.003	45.924	0	0.124	0.135
距医院的交通距离	-0.8874	0.006	-146.864	0	-0.899	-0.876

统计量名称	数值
Omnibus	261.535
Prob(Omnibus)	0
Skew	0.011
Kurtosis	3.918
Durbin-Watson	1.992
Jarque-Bera (JB)	532.933
Prob(JB)	1.88e-116
Cond. No.	47.9
R-squared	0.764
Adj. R-squared	0.764
F-statistic	6806
Prob (F-statistic)	0
AIC	-2360
BIC	-2299

修正回归估计结果表明模型R²值仍然为0.764，说明模型对于金县房价变异性解释能力依然很高，达到76.4%。修正后的F统计量为6806，比逐步回归阶段（nonrobust）的7033下降。根据统计学逻辑可知，该项数值变化是由于稳健标准误对原OLS估计中内生变量异方差造成的显著性虚高进行必要的修正所导致的，直观上体现模型消除扰动项干扰后真实统计显著性的特点。将全变量初始模型录得的4925个数据进行修正后，得到修正模型F统计量有实质性的提升。AIC值和BIC值修正之后有所改善，说明修正后的模型拟合度更好，变量的选择也更合理。

从回归系数可以看出修正模型中所有的保留变量都具有很高的统计显著性。居住面积、海景房、视野、成色、建筑设计水平这五个因素分别与销售价格呈正相关，系数分别为0.1849、0.4697、0.0513、0.0464、0.1299。相反，卧室数的系数是负的，是-0.0083，距医院的交通距离的系数也是负的，是-0.8874。

最后，对修正模型的共线性和残差检验也证明了它统计上是稳健的。

表7：修正模型方差膨胀因子（VIF）检验结果	
Table 7: Variance inflation factor(VIF) of the correct model	
变量	VIF
const	4.4144372061253065
卧室数	1.7446944190424931
居住面积	2.894886100284559
是否海景房	1.0393680266384369
视野	1.0898385481944275
成色	1.0539530899222413
建筑设计水平	2.0347492633835946
距医院的交通距离	1.041662436671749

方差膨胀因子（VIF）检验结果说明，在修正架构中各个变量的VIF值都大于初始模型中各变量的VIF值。除居住面积(2.89)、建筑设计水平(2.03)外，其余变量的VIF值都小于1，并且都远远小于10的临界值。由此可知模型内部的多重共线性问题已经被很好地解决，对应的系数估计的方差变小并且得到了比较稳定的估计值。

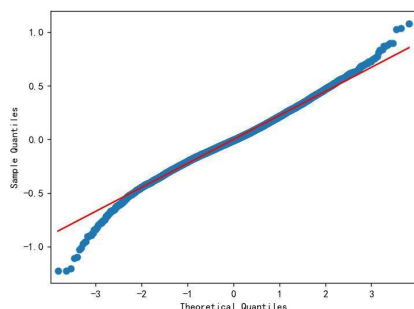


图7: 修正模型残差正态 Q-Q 图

Figure 7: Normal Q-Q plot of the residuals of the corrected model

从观测的残差分布特征可知,修正模型的正态 Q-Q 图中绝大部分样本点都接近于 45 度对角线的轨迹上,说明残差基本上满足正态分布的假设。

五、结论与建议

(一) 实证结果分析

根据逐步回归结果及修正后的模型,可得如下多元回归模型

$$y = 13.4308 - 0.0083x_1 + 0.1849x_2 + 0.4697x_3 + 0.0513x_4 + 0.0464x_5 + 0.1299x_6 - 0.8874x_7$$

以销售价格为 y , 卧室数、居住面积、是否海景房、视野、成色、建筑设计水平、距医院交通距离作为自变量。

房屋的物理属性和设计品质会带来比较大的价格溢价。实证结果表明,居住面积(x_2)、建筑设计水平(x_6)和房屋成色(x_5)的回归系数都是显著正向的,在控制了其他的变量之后,它们都会使房产市场价值得到提高。环境和景观资源的稀缺性也会反映在资本定价体系里。海岸线景观的海景房(x_3)系数为 0.4697,其绝对值较大,说明海岸线景观对房价的影响比较大,视野等级(x_4)正向系数为 0.0513,也体现出了良好的视觉环境有附加价值。

部分反直觉变量的估计结果要联系市场实际来加以具体说明。卧室数(x_1)系数为负(-0.0083),说明在固定总居住面积的基础上,卧室数量增多会使得单间的有效使用空间变小,内部格局变得破碎,居住舒适度和市场吸引力都会降低。距医院交通距离(x_7)的系数为负 0.8874,说明空间距离越大,房产价值越低,很好地证明了金县购房群体对于医疗设施靠近属性的强烈需求,也突出了优质公共服务资源对于房屋价格起着决定性的作用。

(二) 模型预测

完成房价影响因素的静态量化之后,量化模型的进阶价值就是它的预测能力^[7]。根据这样的逻辑,为了进一步检验修正后的模型在样本外的预测效果和泛化程度,本文按照 5% 的比例从金县房屋销售总样本中随机抽取了一个独立的测试集,绘制了模型预测价格与实际成交价格的对比散点图。

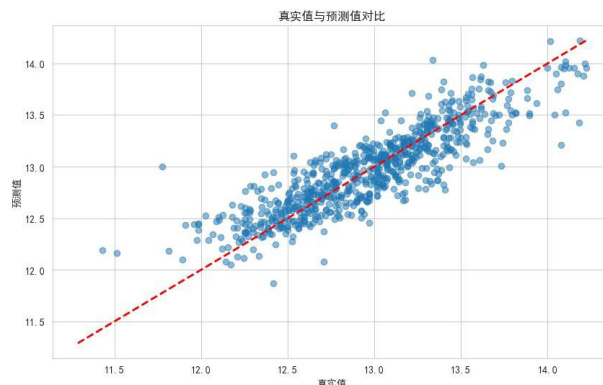


图8: 测试集真实销售价格与预测值对比散点图

Figure 8: Scatter plot showing the comparison between true sale price and predict values in test set.

从分析预测散点图的分布形态可以知道,测试集中绝大部分的观测点都均匀、紧密地分布在 45 度基准线($y=x$)的两侧。该种视觉表现很好地证明了模型的预测结果和实际的房地产市场成交价格十分接近。该结果证明精简后的七变量组合可以很好地捕捉金县房价波动的主要数据生成机制,说明模型对于未被训练过的新的样本仍然有很好的解释力,并没有出现明显的过拟合现象。

从散点图中可以看出,房价分布极值区(图形右上方)的散点越来越向 45 度基准线靠拢,也就是说明该模型对一些高端住宅的预测值偏低。该种局部预测偏差符合特征价格模型典型应用的局限性。真实的房地产定价机制和各种特性之间并不是绝对的纯线性关系,而且市场中还存在着很多无法量化的隐性影响因素,客观上导致了常规的线性回归模型预测值与部分真实极端成交价之间不可避免地产生误差^[8]。超高净值房产的溢价一般是由极端的个人审美偏好、特殊的时空历史人文价值、不可复制的微观地段优势所造成的。这类非标准的、无法用常规的区位变量来解释的隐蔽因素,会打破目前的基础物理以及常规区位变量所造成的解释范围。

(三) 局限性与未来展望

本研究用多重诊断和稳健性修正来提高模型的解释力,但是由于截面数据本身的特性以及传统计量方法的限制,实证框架还存在一些不足。房地产数据具有很强的空间异质性和相关性,相邻区域房价一般都会产生溢出效应。现有的基准多元线性回归模型是以样本独立假设为前提的,不能很好地提取和量化微观地段聚集所造成的空间自相关干扰。

修正模型使用稳健标准误来应对异方差问题,但是这样的处理并没有从根本上改变非同方差的数据产生机制。传统的最小二乘法估计参数有固有的缺陷,部分异常样本很容易对回归系数产生巨大的影响,使输出的参数没有现实的经济意义;面对因变量的变化,传统的回归模型会反应迟钝,不能进行实时的跟踪^[9]。真实的房地产市场常常存在多种变量之间的相互影响,高水平的设计、稀缺的海景资源叠加起来产生的非线性溢价,这样的高阶定价机制很难被现有的线性设定完全包含。

就目前而言，今后的研究可以向空间计量和非参数算法两个方面进行深入。空间维度分析可以采用空间滞后或者误差模型，也可以在基准回归内部嵌入邮政编码等微观区位固定效应，从而剔除掉没有观测到的地段溢价。从方法论上可以使用随机森林等机器学习

算法进行交叉比较，对多变量交互作用下的三维定价曲面做精细的拟合。后续预测模型的构建阶段也可以使用 Stacking 堆叠策略，即采用不同的算法优势互补的方式来提高预测的准确性^[10]，给房地产资产评估和城市空间规划提供更加严谨的量化支撑。

参考文献

[1] 毛晨希, 董可扬, 宋瑾钰. 基于线性回归的房产分析与展示系统设计 [J]. 软件导刊, 2023, 22(07): 104–111.

[2] 张若天. 基于多元线性回归的房价预测模型改进 [J]. 电子制作, 2019, (04): 98–100.

[3] 卢超猛, 马泽众, 韩阳, 郭小强. 基于支持向量回归的房价数据分析 [J]. 华北理工大学学报 (自然科学版), 2021, 43(04): 76–82.

[4] 罗博伟, 洪智勇, 王劲屹. 多元线性回归统计模型在房价预测中的应用 [J]. 计算机时代, 2020, (06): 51–54.

[5] 黄立, 严文婷, 邹美娜, 袁伟琴, 肖奕, 谢斌锋. 基于多元线性回归的赣州市房价影响因素分析 [J]. 内蒙古科技与经济, 2021, (20): 77–78+82.

[6] 钟丽燕, 高淑兰. 多元线性回归模型在房价走势分析与预测中的应用 [J]. 科技创业月刊, 2017, 30(09): 94–96.

[7] 张砚博. 基于多元线性回归分析的西安市房价预测分析 [J]. 西部皮革, 2020, 42(10): 71.

[8] 白招弟, 史佼佼, 裴芳蓉, 闫丽莉, 李媛媛. 甘肃省商品房房价影响因素研究——基于多元线性回归模型的分析 [J]. 广西质量监督导报, 2020, (10): 198–199.

[9] 李正蹊. 基于多元线性回归的房价预测模型 [J]. 中国新通信, 2018, 20(23): 156–158.

[10] 秦艳姣. 基于多元线性回归与随机森林算法的房价预测模型对比研究 [J]. 现代信息科技, 2024, 8(22): 127–131.